



Europäisches
Patentamt

Europäische
Patent Office

Office européen
des brevets

GC 381- P4

PCT/US 98/14529

REC'D 28 SEP 1998
WIPO PCT

Bescheinigung

Certificate

Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

97305227.7

PRIORITY DOCUMENT

SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH RULE 17.1(a) OR (b)

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office
Le Président de l'Office européen des brevets
p.o.

M.B. RIJLING

DEN HAAG, DEN
THE HAGUE, 12/08/98
LA HAYE, LE



Eur päisches
Patentamt

European
Patent Office

Office européen
des brevets

Blatt 2 der Bescheinigung
Sheet 2 of the certificate
Page 2 de l'attestation

Anmeldung Nr.: 97305227.7
Application no.: 97305227.7
Demande n°:

Anmelder:
Applicant(s):
Demandeur(s):
Genencor International B.V.
2600 AP Delft
NETHERLANDS

Anmeldetag:
Date of filing:
Date de dépôt: 15/07/97

Bezeichnung der Erfindung
Title of the invention
Titre de l'invention
Proteases from gram-positive organisms

In Anspruch genommene Prionat(en) / Priority(es) claimed / Priorité(s) revendiquée(s)

Staat	Tag	Aktenzeichen:
State	Date	File no.
Pays	Date	Numéro de dépôt

Internationale Patentklassifikation
International Patent classification
Classification internationale des brevets

/

Am Anmeldetag benannte Vertragsstaaten
Contracting states designated at date of filing: AT/BE/CH/DE/DK/ES/FI/FR/GB/GR/IE/IT/LI/LU/MC/NL/PT/SE
Etats contractants désignés lors du dépôt

Bemerkungen
Remarks
Remarques

PATENT
Docket No. GC 381

PROTEASES FROM GRAM-POSITIVE ORGANISMS

5

FIELD OF THE INVENTION

The present invention relates to cysteine proteases derived from gram-positive microorganisms. The present invention provides nucleic acid and amino acid sequences of cysteine protease 1, 2 and 3 identified in *Bacillus*. The present invention 10 also provides methods for the production of cysteine protease 1, 2 and 3 in host cells as well as the production of heterologous proteins in a host cell having a mutation or deletion of part or all of at least one of the cysteine proteases of the present invention.

BACKGROUND OF THE INVENTION

15 Gram-positive microorganisms, such as members of the group *Bacillus*, have been used for large-scale industrial fermentation due, in part, to their ability to secrete their fermentation products into the culture media. In gram-positive bacteria, secreted proteins are exported across a cell membrane and a cell wall, and then are subsequently released into the external media usually maintaining their native 20 conformation.

25 Various gram-positive microorganisms are known to secrete extracellular and/or intracellular protease at some stage in their life cycles. Many proteases are produced in large quantities for industrial purposes. A negative aspect of the presence of proteases in gram-positive organisms is their contribution to the overall degradation of secreted heterologous or foreign proteins.

The classification of proteases found in microorganisms is based on their catalytic mechanism which results in four groups: the serine proteases; the metalloproteases; cysteine proteases; and aspartic proteases. These categories can be distinguished by their sensitivity to various inhibitors. For example, the serine 30 proteases are inhibited by phenylmethylsulfonylfluoride (PMSF) and diisopropylfluorophosphate (DIFP); the metalloproteases by chelating agents; the cysteine enzymes by iodoacetamide and heavy metals and the aspartic proteases by pepstatin. The serine proteases have alkaline pH optima, the metalloproteases are optimally active around neutrality, and the cysteine and aspartic enzymes have acidic 35 pH optima (*Biotechnology Handbooks, Bacillus*, vol. 2, edited by Harwood, 1989 Plenum Press, New York).

The activity of cysteine protease depends on a catalytic dyad of cysteine and histidine with the order differing between families. The best known family of

cysteine proteases is that of papain having catalytic residues Cys-25 and His-159. Cysteine proteases of the papain family catalyze the hydrolysis of peptide, amide, ester, thiol ester and thione ester bonds. Naturally occurring inhibitors of cysteine proteases of the papain family are those of the cystatin family (Methods in Enzymology, vol. 244, Academic Press, Inc. 1994).

SUMMARY OF THE INVENTION

The present invention relates to the unexpected and surprising discovery of three heretofore unknown or unrecognized cysteine proteases found in *Bacillus subtilis*, designated herein as CP1, CP2 and CP3 having the nucleic acid and amino acid as shown in Figures 1A-1B, Figures 5A-5B and 6A-6B, respectively. The present invention is based, in part, upon the presence of the characteristic cysteine protease amino acid motif GXCWAF found in uncharacterised translated genomic nucleic acid sequences of *Bacillus subtilis*. The present invention is also based in part upon the structural relatedness that CP1 has with the cysteine protease papain specifically with respect to the location of the catalytic histidine/alanine and asparagine/serine residues and the structural relatedness that CP1 has with CP2 and CP3.

The present invention provides isolated polynucleotide and amino acid sequences for CP1, CP2 and CP3. Due to the degeneracy of the genetic code, the present invention encompasses any nucleic acid sequence that encodes the CP1, CP2 and CP3 amino acid sequence shown in the Figures.

The present invention encompasses amino acid variations of *B. subtilis* CP1, CP2 and CP3 amino acids disclosed herein that have proteolytic activity. *B. subtilis* CP1, CP2 and CP3 as well as proteolytically active amino acid variations, thereto have application in cleaning compositions. The present invention also encompasses amino acid variations or derivatives of CP1, CP2, CP3 that do not have the characteristic proteolytic activity as long as the nucleic acid sequences encoding such variations or derivatives would have sufficient 5' and 3' coding regions to be capable of being integrated into a gram-positive organism genome. Such variants would have applications in gram-positive expression systems where it is desirable to delete, mutate, alter or otherwise incapacitate the naturally occurring cysteine protease in order to diminish or delete its proteolytic activity. Such an expression system would have the advantage of allowing for greater yields of recombinant heterologous proteins or polypeptides.

The present invention provides methods for detecting gram positive microorganism homologs of *B. subtilis* CP1, CP2 and CP3 that comprises hybridizing part or all of the nucleic acid encoding *B. subtilis* CP1, CP2 and CP3 with nucleic acid derived from gram-positive organisms, either of genomic or cDNA origin. In one

embodiment, the gram-positive microorganism is selected from the group consisting of *B. licheniformis*, *B. latus*, *B. brevis*, *B. stearothermophilus*, *B. alkalophilus*, *B. amyloliquefaciens*, *B. coagulans*, *B. circulans*, *B. latus* and *Bacillus thuringiensis*.

In yet another aspect, the present invention provides a gram-positive microorganism having a mutation or deletion of part or all of the gene encoding CP1 and/or CP2 and/or CP3, which results in the inactivation of the CP1 and/or CP2 and/or CP3 proteolytic activity, either alone or in combination with mutations in other proteases, such as apr, npr, epr, mpr for example, or other proteases known to those of skill in the art. In one embodiment of the present invention, the gram-positive organism is a member of the genus *Bacillus*. In another embodiment, the *Bacillus* is *Bacillus subtilis*.

The production of desired heterologous proteins or polypeptides in gram-positive microorganisms may be hindered by the presence of one or more proteases which degrade the produced heterologous protein or polypeptide. One advantage of the present invention is that it provides methods and expression systems which can be used to prevent that degradation, thereby enhancing yields of the desired heterologous protein or polypeptide. In another aspect, the gram-positive host having one or more cysteine protease deletions is further genetically engineered to produce a desired protein.

In one embodiment of the present invention, the desired protein is heterologous to the gram-positive host cell. In another embodiment, the desired protein is homologous to the host cell. The present invention encompasses a gram-positive host cell having a deletion or interruption of the nucleic acid encoding the naturally occurring homologous protein, such as a protease, and having nucleic acid encoding the homologous protein re-introduced in a recombinant form. In another embodiment, the host cell produces the homologous protein. Accordingly, the present invention also provides methods and expression systems for reducing degradation of heterologous proteins produced in gram-positive microorganisms. The gram-positive microorganism may be normally sporulating or non-sporulating.

In a further aspect of the present invention, gram-positive CP1, CP2 or CP3 is produced on an industrial fermentation scale in a microbial host expression system. In another aspect, isolated and purified recombinant CP1, CP2 or CP3 is used in compositions of matter intended for cleaning purposes, such as detergents. Accordingly, the present invention provides a cleaning composition comprises one or more of a gram-positive cysteine protease selected from the group consisting of CP1, CP2 and CP3. The cysteine protease may be used alone or in combination with other enzymes and/or mediators or enhancers.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1A-1B shows the DNA and amino acid sequence for CP1 (YJDE).

5 Figure 2 shows an amino acid alignment with papain (accession number papa_carpa.p) with the cysteine protease CP1, designated YJDE. For Figures 2, 3 and 4, the motif GXCWAF has been marked along with the catalytic cysteine and the conserved catalytic histidine/alanine and asparagine/serine residues.

10 Figure 3 shows amino acid alignment of CP1 (YJDE) with CP3 (PMI).

Figure 4 shows the amino acid alignment of CP1 (YJDE) with CP2 (YdhS).

15 Figure 5A-5B shows the amino acid and nucleic acid sequence for CP2 (YdhS).

Figure 6A-6B shows the amino acid and nucleic acid sequence for CP3 (PMI).

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTSDefinitions

20 As used herein, the genus *Bacillus* includes all members known to those of skill in the art, including but not limited to *B. subtilis*, *B. licheniformis*, *B. lentus*, *B. brevis*, *B. stearothermophilus*, *B. alkalophilus*, *B. amyloliquefaciens*, *B. coagulans*, *B. cibulans*, *B. lautus* and *B. thuringiensis*.

25 The present invention encompasses novel CP1, CP2 and CP3 from gram positive organisms. In a preferred embodiment, the gram-positive organisms is a *Bacillus*. In another preferred embodiment, the gram-positive organism is *Bacillus subtilis*. As used herein, "*B. subtilis* CP1, CP2 or CP3" refers to the amino acid sequences shown in Figures. Figures 1A-1B show the amino acid and nucleic acid sequence for CP1 (YJDE); Figures 5A-5B show the amino acid and nucleic acid sequence for CP2 (YDHS); and Figures 6A-6B show the amino acid and nucleic acid 30 sequences for CP3 (PMI). The present invention encompasses amino acid variations of the amino acid sequences disclosed in Figures 1A-1B and 5A-5B and 6A-6B that have proteolytic activity. Such proteolytic amino acid variants can be used in cleaning compositions. The present invention also encompasses *B. subtilis* amino acid variations or derivatives that are not proteolytically active. DNA encoding such 35 variants can be used in methods designed to delete or mutate the naturally occurring host cell CP1, CP or CP3.

As used herein, "nucleic acid" refers to a nucleotide or polynucleotide sequence, and fragments or portions thereof, and to DNA or RNA of genomic or

synthetic origin which may be double-stranded or single-stranded, whether representing the sense or antisense strand. As used herein "amino acid" refers to peptide or protein sequences or portions thereof. A "polynucleotide homolog" as used herein refers to a gram-positive microorganism polynucleotide that has at least 5 80%, at least 90% and at least 95% identity to *B. subtilis* CP1, CP2 or CP3, or which is capable of hybridizing to *B. subtilis* CP1, CP2 or CP3 under conditions of high stringency and which encodes an amino acid sequence having cysteine protease activity.

10 The terms "isolated" or "purified" as used herein refer to a nucleic acid or amino acid that is removed from at least one component with which it is naturally associated.

15 As used herein, the term "heterologous protein" refers to a protein or polypeptide that does not naturally occur in a gram-positive host cell. Examples of heterologous proteins include enzymes such as hydrolases including proteases, cellulases, amylases, carbohydrases, and lipases; isomerases such as racemases, epimerases, tautomerases, or mutases; transferases, kinases and phosphatases. The heterologous gene may encode therapeutically significant proteins or peptides, such as 20 growth factors, cytokines, ligands, receptors and inhibitors, as well as vaccines and antibodies. The gene may encode commercially important industrial proteins or peptides, such as proteases, carbohydrases such as amylases and glucoamylases, cellulases, oxidases and lipases. The gene of interest may be a naturally occurring gene, a mutated gene or a synthetic gene.

25 The term "homologous protein" refers to a protein or polypeptide native or naturally occurring in a gram-positive host cell. The invention includes host cells producing the homologous protein via recombinant DNA technology. The present invention encompasses a gram-positive host cell having a deletion or interruption of the nucleic acid encoding the naturally occurring homologous protein, such as a protease, and having nucleic acid encoding the homologous protein re-introduced in a recombinant form. In another embodiment, the host cell produces the homologous protein.

30 As used herein, the term "overexpressing" when referring to the production of a protein in a host cell means that the protein is produced in greater amounts than its production in its naturally occurring environment.

35 As used herein, the phrase "proteolytic activity" refers to a protein that is able to hydrolyze a peptide bond. Enzymes having proteolytic activity are described in Enzyme Nomenclature, 1992, edited Webb Academic Press, Inc.

Detailed Description of the Preferred Embodiments

The unexpected discovery of the cysteine proteases CP1, CP2 and CP3 in *B. subtilis* provides a basis for producing host cells, expression methods and systems which can be used to prevent the degradation of recombinantly produced heterologous 5 proteins. In a preferred embodiment, the host cell is a gram-positive host cell that has a deletion or mutation in the naturally occurring cysteine protease said mutation resulting in deletion or inactivation of the production by the host cell of the proteolytic cysteine protease gene product. The host cell may additionally be genetically engineered to produce a desired protein or polypeptide.

10 It may also be desired to genetically engineer host cells of any type to produce a gram-positive cysteine protease. Such host cells are used in large scale fermentation to produce large quantities of the cysteine protease which may be isolated or purified and used in cleaning products, such as detergents.

15 I. Cysteine Protease Sequences

The CP1, CP2 and CP3 polynucleotides having the sequences as shown in Figures 1A-1B, 5A-5B and 6A-6B, respectively, encode the *Bacillus subtilis* cysteine 20 proteases CP1, CP2 and CP3. As will be understood by the skilled artisan, due to the degeneracy of the genetic code, a variety of polynucleotides can encode the *Bacillus subtilis* CP1, CP2 and CP3. The present invention encompasses all such polynucleotides.

The present invention encompasses CP1, CP2 and CP3 polynucleotide homologs encoding gram-positive microorganism cysteine proteases CP1, CP2 and CP3, respectively, which have at least 80%, or at least 90% or at least 95% identity to *B. subtilis* CP1, CP2 and CP3 as long as the homolog encodes a protein that has proteolytic activity.

Gram-positive polynucleotide homologs of *B. subtilis* CP1, CP2 or CP3 may be obtained by standard procedures known in the art from, for example, cloned DNA (e.g., a DNA "library": genomic DNA libraries, by chemical synthesis once identified, by 25 cDNA cloning, or by the cloning of genomic DNA, or fragments thereof, purified from a desired cell. See, for example, Sambrook *et al.*, 1989, Molecular Cloning, A Laboratory Manual, 2d Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York; Glover, D.M. (ed.), 1985, DNA Cloning: A Practical Approach, MRL Press, Ltd., Oxford, U.K. Vol. 1, 2). A preferred source is from genomic DNA. Nucleic acid 30 sequences derived from genomic DNA may contain regulatory regions in addition to coding regions. Whatever the source, the isolated CP1, CP2 or CP3 gene should be molecularly cloned into a suitable vector for propagation of the gene.

In the molecular cloning of the gene from genomic DNA, DNA fragments are generated, some of which will encode the desired gene. The DNA may be cleaved at specific sites using various restriction enzymes. Alternatively, one may use DNase in the presence of manganese to fragment the DNA, or the DNA can be physically sheared, 5 as for example, by sonication. The linear DNA fragments can then be separated according to size by standard techniques, including but not limited to, agarose and polyacrylamide gel electrophoresis and column chromatography.

Once the DNA fragments are generated, identification of the specific DNA fragment containing the CP1, CP2 or CP3 may be accomplished in a number of ways. 10 For example, a *B. subtilis* CP1, CP2 or CP3 gene of the present invention or its specific RNA, or a fragment thereof, such as a probe or primer, may be isolated and labeled and then used in hybridization assays to detect a gram-positive CP1, CP2 or CP3 gene. (Benton, W. and Davis, R., 1977, Science 196:180; Grunstein, M. And Hogness, D., 1975, Proc. Natl. Acad. Sci. USA 72:3961). Those DNA fragments 15 sharing substantial sequence similarity to the probe will hybridize under stringent conditions.

Accordingly, the present invention provides a method for the detection of gram-positive CP1, CP2 and CP3 polynucleotide homologs which comprises hybridizing part or all of a nucleic acid sequence of *B. subtilis* CP1, CP2 and CP3 20 with gram-positive microorganism nucleic acid of either genomic or cDNA origin.

Also included within the scope of the present invention are gram-positive microorganism polynucleotide sequences that are capable of hybridizing to the nucleotide sequence of *B. subtilis* CP1, CP2 or CP3 under conditions of intermediate to maximal stringency. Hybridization conditions are based on the melting 25 temperature (T_m) of the nucleic acid binding complex, as taught in Berger and Kimmel (1987, Guide to Molecular Cloning Techniques, Methods in Enzymology, Vol 152, Academic Press, San Diego CA) incorporated herein by reference, and confer a defined "stringency" as explained below.

"Maximum stringency" typically occurs at about T_m-5°C (5°C below the T_m 30 of the probe); "high stringency" at about 5°C to 10°C below T_m; "intermediate stringency" at about 10°C to 20°C below T_m; and "low stringency" at about 20°C to 25°C below T_m. As will be understood by those of skill in the art, a maximum stringency hybridization can be used to identify or detect identical polynucleotide sequences while an intermediate or low stringency hybridization can be used to 35 identify or detect polynucleotide sequence homologs.

The term "hybridization" as used herein shall include "the process by which a strand of nucleic acid joins with a complementary strand through base pairing" (Coombs J (1994) Dictionary of Biotechnology, Stockton Press, New York NY).

5 The process of amplification as carried out in polymerase chain reaction (PCR) technologies is described in Dieffenbach CW and GS Dveksler (1995, PCR Primer, a Laboratory Manual, Cold Spring Harbor Press, Plainview NY). A nucleic acid sequence of at least about 10 nucleotides and as many as about 60 nucleotides from *B. subtilis* CP1, CP2 or CP3 preferably about 12 to 30 nucleotides, and more preferably about 20-25 nucleotides can be used as a probe or PCR primer.

10 The *B. subtilis* amino acid sequences CP1, CP2 and CP3 (shown in Figures 2, 4 and 3, respectively) were identified via a FASTA search of *Bacillus subtilis* genomic nucleic acid sequences. *B. subtilis* CP1 (YJDE) was identified by its structural homology to the cysteine protease papain having the sequence designated "papa_carpa.p". As shown in Figure 2, YJDE has the motif GXCWAF as well as the 15 conserved catalytic residues His/Ala and Asn/Ser. CP2 (YdHS) and CP3 (PMI) were identified upon their structural homology to CP1 (YJDE). The presence of GXCWAF as well as residues His/Ala and Asn/Ser is noted in Figures 3 and 4. CP3 (PMI) was previously characterized as a possible phosphomannose isomerase, (Noramata). There has been no previous characterization of CP3 as a cysteine protease.

20

II. Expression Systems

The present invention provides host cells, expression methods and systems for the enhanced production and secretion of desired heterologous or homologous proteins in gram-positive microorganisms. In one embodiment, a host cell is 25 genetically engineered to have a deletion or mutation in the gene encoding a gram-positive CP1, CP2 or CP3 such that the respective activity is deleted. In another embodiment of the present invention, a gram-positive microorganism is genetically engineered to produce a cysteine protease of the present invention.

Inactivation of a gram-positive cysteine protease in a host cell

30 Producing an expression host cell incapable of producing the naturally occurring cysteine protease necessitates the replacement and/or inactivation of the naturally occurring gene from the genome of the host cell. In a preferred embodiment, the mutation is a non-reverting mutation.

One method for mutating nucleic acid encoding a gram-positive cysteine protease is to clone the nucleic acid or part thereof, modify the nucleic acid by site directed mutagenesis and reintroduce the mutated nucleic acid into the cell on a plasmid. By homologous recombination, the mutated gene may be introduced into the 5 chromosome. In the parent host cell, the result is that the naturally occurring nucleic acid and the mutated nucleic acid are located in tandem on the chromosome. After a second recombination, the modified sequence is left in the chromosome having thereby effectively introduced the mutation into the chromosomal gene for progeny of the parent host cell.

10 Another method for inactivating the cysteine protease proteolytic activity is through deleting the chromosomal gene copy. In a preferred embodiment, the entire gene is deleted, the deletion occurring in such a way as to make reversion impossible. In another preferred embodiment, a partial deletion is produced, provided that the nucleic acid sequence left in the chromosome is too short for homologous 15 recombination with a plasmid encoded cysteine protease gene. In another preferred embodiment, nucleic acid encoding the catalytic amino acid residues are deleted.

20 Deletion of the naturally occurring gram-positive microorganism cysteine protease can be carried out as follows. A cysteine protease gene including its 5' and 3' regions is isolated and inserted into a cloning vector. The coding region of the cysteine protease gene is deleted from the vector *in vitro*, leaving behind a sufficient amount of the 5' and 3' flanking sequences to provide for homologous recombination 25 with the naturally occurring gene in the parent host cell. The vector is then transformed into the gram-positive host cell. The vector integrates into the chromosome via homologous recombination in the flanking regions. This method leads to a gram-positive strain in which the protease gene has been deleted.

30 The vector used in an integration method is preferably a plasmid. A selectable marker may be included to allow for ease of identification of desired recombinant microorganisms. Additionally, as will be appreciated by one of skill in the art, the vector is preferably one which can be selectively integrated into the chromosome. This can be achieved by introducing an inducible origin of replication, for example, a 35 temperature sensitive origin into the plasmid. By growing the transformants at a temperature to which the origin of replication is sensitive, the replication function of the plasmid is inactivated, thereby providing a means for selection of chromosomal integrants. Integrants may be selected for growth at high temperatures in the presence of the selectable marker, such as an antibiotic. Integration mechanisms are described in WO 88/06623.

Integration by the Campbell-type mechanism can take place in the 5' flanking region of the protease gene, resulting in a protease positive strain carrying the entire

plasmid vector in the chromosome in the cysteine protease locus. Since illegitimate recombination will give different results it will be necessary to determine whether the complete gene has been deleted, such as through nucleic acid sequencing or restriction maps.

5 Another method of inactivating the naturally occurring cysteine protease gene is to mutagenize the chromosomal gene copy by transforming a gram-positive microorganism with oligonucleotides which are mutagenic. Alternatively, the chromosomal cysteine protease gene can be replaced with a mutant gene by homologous recombination.

10 The present invention encompasses host cells having additional protease deletions or mutations, such as deletions or mutations in apr, npr, epr, mpr and others known to those of skill in the art.

15 One assay for the detection of mutants involves growing the *Bacillus* host cell on medium containing a protease substrate and measuring the appearance or lack thereof, of a zone of clearing or halo around the colonies. Host cells which have an inactive protease will exhibit little or no halo around the colonies.

III. Production of Cysteine Protease

20 For production of cysteine protease in a host cell, an expression vector comprising at least one copy of nucleic acid encoding a gram-positive microorganism CP1, CP2 or CP3, and preferably comprising multiple copies, is transformed into the host cell under conditions suitable for expression of the cysteine protease. In accordance with the present invention, polynucleotides which encode a gram-positive microorganism CP1, CP2 or CP3, or fragments thereof, or fusion proteins or 25 polynucleotide homolog sequences that encode amino acid variants of *B. subtilis* CP1, CP2 or CP3, may be used to generate recombinant DNA molecules that direct their expression in host cells. In a preferred embodiment, the gram-positive host cell belongs to the genus *Bacillus*. In another preferred embodiment, the gram positive host cell is *B. subtilis*.

30 As will be understood by those of skill in the art, it may be advantageous to produce polynucleotide sequences possessing non-naturally occurring codons. Codons preferred by a particular gram-positive host cell (Murray E et al (1989) Nuc Acids Res 17:477-508) can be selected, for example, to increase the rate of expression or to produce recombinant RNA transcripts having desirable properties, such as a longer 35 half-life, than transcripts produced from naturally occurring sequence.

Altered CP1, CP2 or CP3 polynucleotide sequences which may be used in accordance with the invention include deletions, insertions or substitutions of different nucleotide residues resulting in a polynucleotide that encodes the same or a

functionally equivalent CP1, CP2 or CP3 homolog, respectively. As used herein a "deletion" is defined as a change in either nucleotide or amino acid sequence in which one or more nucleotides or amino acid residues, respectively, are absent.

As used herein an "insertion" or "addition" is that change in a nucleotide or 5 amino acid sequence which has resulted in the addition of one or more nucleotides or amino acid residues, respectively, as compared to the naturally occurring CP1, CP3 or CP3.

As used herein "substitution" results from the replacement of one or more nucleotides or amino acids by different nucleotides or amino acids, respectively.

10 The encoded protein may also show deletions, insertions or substitutions of amino acid residues which produce a silent change and result in a functionally CP1, CP2 or CP3 variant. Deliberate amino acid substitutions may be made on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues as long as the variant retains the ability to modulate 15 secretion. For example, negatively charged amino acids include aspartic acid and glutamic acid; positively charged amino acids include lysine and arginine; and amino acids with uncharged polar head groups having similar hydrophilicity values include leucine, isoleucine, valine; glycine, alanine; asparagine, glutamine; serine, threonine, phenylalanine, and tyrosine.

20 The CP1, CP2 or CP3 polynucleotides of the present invention may be engineered in order to modify the cloning, processing and/or expression of the gene product. For example, mutations may be introduced using techniques which are well known in the art, e.g. site-directed mutagenesis to insert new restriction sites, to alter glycosylation patterns or to change codon preference, for example.

25 In one embodiment of the present invention, a gram-positive microorganism CP1, CP2 or CP3 polynucleotide may be ligated to a heterologous sequence to encode a fusion protein. A fusion protein may also be engineered to contain a cleavage site located between the cysteine protease nucleotide sequence and the heterologous protein sequence, so that the cysteine protease may be cleaved and purified away from 30 the heterologous protein.

IV. Vector Sequences

Expression vectors used in expressing the cysteine proteases of the present invention in gram-positive microorganisms comprise at least one promoter associated 35 with a cysteine protease selected from the group consisting of CP1, CP2 and CP3, which promoter is functional in the host cell. In one embodiment of the present invention, the promoter is the wild-type promoter for the selected cysteine protease and in another embodiment of the present invention, the promoter is heterologous to

the cysteine protease, but still functional in the host cell. In one preferred embodiment of the present invention, nucleic acid encoding the cysteine protease is stably integrated into the microorganism genome.

In a preferred embodiment, the expression vector contains a multiple cloning site cassette which preferably comprises at least one restriction endonuclease site unique to the vector, to facilitate ease of nucleic acid manipulation. In a preferred embodiment, the vector also comprises one or more selectable markers. As used herein, the term **selectable marker** refers to a gene capable of expression in the gram-positive host which allows for ease of selection of those hosts containing the vector. Examples of such selectable markers include but are not limited to antibiotics, such as, erythromycin, actinomycin, chloramphenicol and tetracycline.

V. Transformation

A variety of host cells can be used for the production of CP1, CP2 and CP3 including bacterial, fungal, mammalian and insects cells. General transformation procedures are taught in *Current Protocols In Molecular Biology* (vol. 1, edited by Ausubel et al., John Wiley & Sons, Inc. 1987, Chapter 9) and include calcium phosphate methods, transformation using DEAE-Dextran and electroporation. Plant transformation methods are taught in Rodriguez (WO 95/14099, published 26 May 1995).

In a preferred embodiment, the host cell is a gram-positive microorganism and in another preferred embodiment, the host cell is *Bacillus*. In one embodiment of the present invention, nucleic acid encoding one or more cysteine protease(s) of the present invention is introduced into a host cell via an expression vector capable of replicating within the *Bacillus* host cell. Suitable replicating plasmids for *Bacillus* are described in *Molecular Biological Methods for Bacillus*, Ed. Harwood and Cutting, John Wiley & Sons, 1990, hereby expressly incorporated by reference; see chapter 3 on plasmids. Suitable replicating plasmids for *B. subtilis* are listed on page 92.

In another embodiment, nucleic acid encoding a cysteine protease(s) of the present invention is stably integrated into the microorganism genome. Preferred host cells are gram-positive host cells. Another preferred host is *Bacillus*. Another preferred host is *Bacillus subtilis*. Several strategies have been described in the literature for the direct cloning of DNA in *Bacillus*. Plasmid marker rescue transformation involves the uptake of a donor plasmid by competent cells carrying a partially homologous resident plasmid (Contente et al., Plasmid 2:555-571 (1979); Haima et al., Mol. Gen. Genet. 223:185-191 (1990); Weinrauch et al., J. Bacteriol. 154(3):1077-1087 (1983); and Weinrauch et al., J. Bacteriol. 169(3):1205-1211

(1987)). The incoming donor plasmid recombines with the homologous region of the resident "helper" plasmid in a process that mimics chromosomal transformation.

Transformation by protoplast transformation is described for *B. subtilis* in Chang and Cohen, (1979) Mol. Gen. Genet. 168:111-115; for *B. megaterium* in 5 Vorobjeva et al., (1980) FEMS Microbiol. Letters 7:261-263; for *B. amyloliquefaciens* in Smith et al., (1986) Appl. and Env. Microbiol. 51:634; for *B. thuringiensis* in Fisher et al., (1981) Arch. Microbiol. 139:213-217; for *B. sphaericus* in McDonald (1984) J. Gen. Microbiol. 130:203; and *B. larvae* in Bakhet et al., (1985) 49:577. Mann et al., (1986, Current Microbiol. 13:131-135) 10 report on transformation of *Bacillus* protoplasts and Holubova, (1985) Folia Microbiol. 30:97) disclose methods for introducing DNA into protoplasts using DNA containing liposomes.

VI. Identification of Transformants

15 Whether a host cell has been transformed with a mutated or a naturally occurring gene encoding a gram-positive CP1, CP2 or CP3, detection of the presence/absence of marker gene expression can suggest whether the gene of interest is present. However, its expression should be confirmed. For example, if the nucleic acid encoding a cysteine protease is inserted within a marker gene sequence, 20 recombinant cells containing the insert can be identified by the absence of marker gene function. Alternatively, a marker gene can be placed in tandem with nucleic acid encoding the cysteine protease under the control of a single promoter. Expression of the marker gene in response to induction or selection usually indicates expression of the cysteine protease as well.

25 Alternatively, host cells which contain the coding sequence for a cysteine protease and express the protein may be identified by a variety of procedures known to those of skill in the art. These procedures include, but are not limited to, DNA-DNA or DNA-RNA hybridization and protein bioassay or immunoassay techniques which include membrane-based, solution-based, or chip-based technologies for the 30 detection and/or quantification of the nucleic acid or protein.

The presence of the cysteine polynucleotide sequence can be detected by DNA-DNA or DNA-RNA hybridization or amplification using probes, portions or fragments of *B. subtilis* CP1, CP2 or CP3.

35 VII Assay of Protease Activity

There are various assays known to those of skill in the art for detecting and measuring protease activity. There are assays based upon the release of acid-soluble peptides from casein or hemoglobin measured as absorbance at 280 nm or

colorimetrically using the Folin method (Bergmeyer, et al., 1984, Methods of Enzymatic Analysis vol. 5, Peptidases, Proteinases and their Inhibitors, Verlag Chemie, Weinheim). Other assays involve the solubilization of chromogenic substrates (Ward, 1983, Proteinases, in Microbial Enzymes and Biotechnology (W.M. Fogarty, ed.), Applied Science, London, pp. 251-317).

VIII Secretion of Recombinant Proteins

Means for determining the levels of secretion of a heterologous or homologous protein in a gram-positive host cell and detecting secreted proteins include, using either polyclonal or monoclonal antibodies specific for the protein. Examples include enzyme-linked immunosorbent assay (ELISA), radioimmunoassay (RIA) and fluorescent activated cell sorting (FACS). These and other assays are described, among other places, in Hampton R et al (1990, Serological Methods, a Laboratory Manual, APS Press, St Paul MN) and Maddox DE et al (1983, J Exp Med 158:1211).

A wide variety of labels and conjugation techniques are known by those skilled in the art and can be used in various nucleic acid and amino acid assays. Means for producing labeled hybridization or PCR probes for detecting specific polynucleotide sequences include oligolabeling, nick translation, end-labeling or PCR amplification using a labeled nucleotide. Alternatively, the nucleotide sequence, or any portion of it, may be cloned into a vector for the production of an mRNA probe. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes *in vitro* by addition of an appropriate RNA polymerase such as T7, T3 or SP6 and labeled nucleotides.

A number of companies such as Pharmacia Biotech (Piscataway NJ), Promega (Madison WI), and US Biochemical Corp (Cleveland OH) supply commercial kits and protocols for these procedures. Suitable reporter molecules or labels include those radionuclides, enzymes, fluorescent, chemiluminescent, or chromogenic agents as well as substrates, cofactors, inhibitors, magnetic particles and the like. Patents teaching the use of such labels include US Patents 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149 and 4,366,241. Also, recombinant immunoglobulins may be produced as shown in US Patent No. 4,816,567 and incorporated herein by reference.

IX Purification of Proteins

Gram positive host cells transformed with polynucleotide sequences encoding heterologous or homologous protein may be cultured under conditions suitable for the expression and recovery of the encoded protein from cell culture. The protein produced by a recombinant gram-positive host cell comprising a mutation or deletion

of the cysteine protease activity will be secreted into the culture media. Other recombinant constructions may join the heterologous or homologous polynucleotide sequences to nucleotide sequence encoding a polypeptide domain which will facilitate purification of soluble proteins (Kroll DJ et al (1993) DNA Cell Biol 12:441-53).

5 Such purification facilitating domains include, but are not limited to, metal chelating peptides such as histidine-tryptophan modules that allow purification on immobilized metals (Porath J (1992) Protein Expr Purif 3:263-281), protein A domains that allow purification on immobilized immunoglobulin, and the domain utilized in the FLAGS extension/affinity purification system (Immunex Corp, Seattle 10 WA). The inclusion of a cleavable linker sequence such as Factor XA or enterokinase (Invitrogen, San Diego CA) between the purification domain and the heterologous protein can be used to facilitate purification.

X Uses of The Present Invention

15 CP1, CP2 and CP3 and Genetically Engineered Host Cells

The present invention provides genetically engineered host cells comprising preferably non-revertable mutations or deletions in the naturally occurring gene encoding CP1, CP2 or CP3 such that the proteolytic activity is diminished or deleted altogether. The host cell may contain additional protease deletions, such as deletions 20 of the mature subtilisin protease and/or mature neutral protease disclosed in United States Patent No. 5,264,366.

In a preferred embodiment, the host cell is further genetically engineered to produce a desired protein or polypeptide. In a preferred embodiment the host cell is a *Bacillus*. In another preferred embodiment, the host cell is a *Bacillus subtilis*.

25 In an alternative embodiment, a host cell is genetically engineered to produce a gram-positive CP1, CP2 or CP3. In a preferred embodiment, the host cell is grown under large scale fermentation conditions, the CP1, CP2 or CP3 is isolated and/or purified and used in cleaning compositions such as detergents. WO 95/10615 discloses detergent formulation.

30 CP1, CP2 and CP3 Polynucleotides

A *B. subtilis* polynucleotide, or any part thereof, provides the basis for detecting the presence of gram-positive microorganism polynucleotide homologs through hybridization techniques and PCR technology.

Accordingly, one aspect of the present invention is to provide for nucleic acid 35 hybridization and PCR probes which can be used to detect polynucleotide sequences, including genomic and cDNA sequences, encoding gram-positive CP1, CP2 or CP3 or portions thereof.

The manner and method of carrying out the present invention may be more fully understood by those of skill in the art by reference to the following examples, which examples are not intended in any manner to limit the scope of the present invention or of the claims directed thereto

5

Example I
Preparation of a Genomic library

The following example illustrates the preparation of a *Bacillus* genomic library.

10 Genomic DNA from *Bacillus* cells is prepared as taught in Current Protocols In Molecular Biology vol. 1, edited by Ausubel et al., John Wiley & Sons, Inc. 1987, chapter 2. 4.1. Generally, *Bacillus* cells from a saturated liquid culture are lysed and the proteins removed by digestion with proteinase K. Cell wall debris, polysaccharides, and remaining proteins are removed by selective precipitation with 15 CTAB, and high molecular weight genomic DNA is recovered from the resulting supernatant by isopropanol precipitation. If exceptionally clean genomic DNA is desired, an additional step of purifying the *Bacillus* genomic DNA on a cesium chloride gradient is added.

20 After obtaining purified genomic DNA, the DNA is subjected to Sau3A digestion. Sau3A recognizes the 4 base pair site GATC and generates fragments compatible with several convenient phage lambda and cosmid vectors. The DNA is subjected to partial digestion to increase the chance of obtaining random fragments.

25 The partially digested *Bacillus* genomic DNA is subjected to size fractionation on a 1% agarose gel prior to cloning into a vector. Alternatively, size fractionation on a sucrose gradient can be used. The genomic DNA obtained from the size fractionation step is purified away from the agarose and ligated into a cloning vector appropriate for use in a host cell and transformed into the host cell.

Example II

30 **Detection of gram-positive microorganisms**

The following example describes the detection of gram-positive microorganism CP1. The same procedures can be used to detect CP2 and CP3.

35 DNA derived from a gram-positive microorganism is prepared according to the methods disclosed in Current Protocols in Molecular Biology, Chap. 2 or 3. The nucleic acid is subjected to hybridization and/or PCR amplification with a probe or primer derived from CP1. A preferred probe comprises the nucleic acid section containing the conserved motif GXCWAF.

5 The nucleic acid probe is labeled by combining 50 pmol of the nucleic acid and 250 mCi of [gamma ³²P] adenosine triphosphate (Amersham, Chicago IL) and T4 polynucleotide kinase (DuPont NEN[®], Boston MA). The labeled probe is purified with Sephadex G-25 super fine resin column (Pharmacia). A portion containing 10⁷ counts per minute of each is used in a typical membrane based hybridization analysis of nucleic acid sample of either genomic or cDNA origin.

10 The DNA sample which has been subjected to restriction endonuclease digestion is fractionated on a 0.7 percent agarose gel and transferred to nylon membranes (Nytran Plus, Schleicher & Schuell, Durham NH). Hybridization is carried out for 16 hours at 40 degrees C. To remove nonspecific signals, blots are sequentially washed at room temperature under increasingly stringent conditions up to 0.1 x saline sodium citrate and 0.5% sodium dodecyl sulfate. The blots are exposed to film for several hours, the film developed and hybridization patterns are compared visually to detect polynucleotide homologs of *B. subtilis* CP1. The homologs are subjected to confirmatory nucleic acid sequencing. Methods for nucleic acid sequencing are well known in the art. Conventional enzymatic methods employ DNA polymerase Klenow fragment, SEQUENASE[®] (US Biochemical Corp, Cleveland, OH) or Taq polymerase to extend DNA chains from an oligonucleotide primer annealed to the DNA template of interest.

20 Various other examples and modifications of the foregoing description and examples will be apparent to a person skilled in the art after reading the disclosure without departing from the spirit and scope of the invention, and it is intended that all such examples or modifications be included within the scope of the appended claims. All publications and patents referenced herein are hereby incorporated by reference in their entirety.

CLAIMS

We claim:

1. An isolated polynucleotide encoding CP1 from a gram positive microorganism.
- 5 2. The polynucleotide of Claim 1 wherein CP1 has the amino acid sequence shown in Figures 1A-1B.
3. An isolated CP1 encoding nucleic acid having the nucleic acid sequence as shown in Figure 1.
- 10 4. An isolated CP1 from a gram-positive microorganism.
5. The isolated CP1 of Claim 4 having the amino acid sequence as shown in Figures 1A-1B.
- 15 6. An isolated polynucleotide encoding CP2 from a gram positive microorganism.
7. The polynucleotide of Claim 6 wherein CP2 has the amino acid sequence shown in Figures 5A-5B.
- 20 8. The isolated CP2 encoding nucleic acid having the sequence as shown in Figures 5A-5B.
9. An isolated CP2 from a gram-positive microorganism.
- 25 10. The isolated CP2 of Claim 9 having the amino acid sequence as shown in Figures 5A-5B.
11. A gram-positive microorganism having a mutation or deletion of part or all of the 30 gene encoding CP1 said mutation or deletion resulting in the inactivation of the CP1 proteolytic activity.
12. A gram-positive microorganism having a mutation or deletion of part or all of the 35 gene encoding CP2 said mutation or deletion resulting in the inactivation of the CP2 proteolytic activity.

13. A gram-positive microorganism having a mutation or deletion of part or all of the gene encoding CP3 said mutation or deletion resulting in the inactivation of the CP3 proteolytic activity.
14. The gram-positive microorganism according to Claims 11, 12 or 13 that is a member of the family *Bacillus*.
15. The microorganism according to Claim 14 wherein the member is selected from the group consisting of *B. licheniformis*, *B. lentus*, *B. brevis*, *B. stearothermophilus*, *B. alkalophilus*, *B. amyloliquefaciens*, *B. coagulans*, *B. circulans*, *B. lautus* and *Bacillus thuringiensis*.
16. The microorganism of Claim 11, 12 or 13 wherein said microorganism is capable of expressing a heterologous protein.
17. The host cell of Claim 16 wherein said heterologous protein is selected from the group consisting of hormone, enzyme, growth factor and cytokine.
18. The host cell of Claim 17 wherein said heterologous protein is an enzyme.
19. The host cell of Claim 15 wherein said enzyme is selected from the group consisting of a proteases, carbohydrases, and lipases; isomerases such as racemases, epimerases, tautomerases, or mutases; transferases, kinases and phosphatases.
20. A cleaning composition comprising a cysteine protease selected from the group consisting of CP1, CP2 and CP3.
21. An expression vector comprising nucleic acid encoding a cysteine protease selected from the group consisting of CP1, CP2 and CP3.
22. A host cell comprising an expression vector according to Claim 21.

ABSTRACT

The present invention relates to the identification of novel cysteine proteases in Gram-positive microorganisms. The present invention provides the nucleic acid and amino acid sequences for the *Bacillus subtilis* cysteine proteases CP1, CP2 and CP3. 5 The present invention also provides host cells having a mutation or deletion of part or all of the gene encoding CP1, CP2 or CP3. The present invention also provides host cells further comprising nucleic acid encoding desired heterologous proteins such as enzymes. The present invention also provides a cleaning composition comprising a cysteine protease of the present invention.

10 - - - - - 30
 atgacgactgaaccgttattttcaaggctgtttcaaagaaaagaatt
 M T T E P L F F K P V F K E R I

 50 70 90
 tggggcgggaccgccttagctgatggctataaccattccgtcaca
 W G G T A L A D F G Y T I P S Q

 110 130
 cgaacaggggagtgctgggctttgccgcgcataaaatggtcaaagc
 R T G E C W A F A A H Q N G Q S

 150 170 190
 gttgtcaaaacggaatgtataaggggttcacgcctcagcgaattatgg
 V V Q N G M Y K G F T L S E L W

 210 230
 gaacatcacagacatttattcggacagcttgaaggggaccgttccct
 E H H R H L F G Q L E G D R F P

 250 270 2
 ctgctacaaaatatttagatgctgaccaggacttatctgtcaggtg
 L L T K I L D A D Q D L S V Q V

 90 310 330
 catccgaatgatgaatatgccaacatacatgaaaacggtgagcttgg
 H P N D E Y A N I H E N G E L G

 350 370
 aaaacagaatgctggtacatttattgattgccaaaaagatgccgagatt
 K T E C W Y I I D C Q K D A E I

 390 410 430
 atttatggccacaatgcaacaacaaggagaagaactaactaccatgata
 I Y G H N A T T K E E L T T M I

 450 470
 gagcgtggagaatggatgagcttgcgcgtgtaaaggtaagccg
 E R G E W D E L L R R V K V K P

 490 510 5
 ggggatttttctatgtgccaagcggtactgttcatgcgtggaaaa
 G D F F Y V P S G T V H A I G K

 30 550 570
 ggaattcttgcttggagacgcagcagaactcagacacaacctacaga
 G I L A L E T Q Q N S D T T Y R

 590 610
 ttatatgattatgaccgaaaagatgcagaaggcaagctgcgcgagctt
 L Y D Y D R K D A E G K L R E L

 630 650 670
 catctgaaaaagagcattgaagtgatagaggcccgtctattccagaa
 H L K K S I E V I E V P S I P E

 690 710

GC 381 Figure 1A

77
cggcatacagttcaccatgaacaaattgaggattgcttacaacgaca

R H T V H H E Q I E D L L T T T

730 750 7
ttgattgaatgcgcctactttcggtgggaaatgaaacttatcagga
L I E C A Y F S V G K W N L S G

70 790 810
tcagcaagcttaaaggcagcaaaaaccatccttcttatcagtgatt
S A S L K Q Q K P F L L I S V I

830 850
gaaggggagggccgtatgatctctggtagtatgtctatcccttcaaa
E G E G R M I S G E Y V Y P F K

870 890 910
aaaggagatcatatgttgctgccttacggtcttggagaattaaactc
K G D H M L L P Y G L G E F K L

930
gaaggatatgcagaatgtatcgctccatctg
E G Y A E C I V S H L

GC 381 Figure 1B

SCORES Init1: 48 Initn: 48 Opt: 79 z-score: 94.7 E():
7.2

Smith-Waterman score: 79; 21.9% identity in 155 aa overlap

130 140 150 160 170 180
papa_carpa.p VLNDGDVNIP~~EY~~YDWRQKGAVTPVKNQGSGC~~W~~AFSAV~~T~~IEGI~~K~~IRTGNLNEYSE
QE

YJDE PLFFKPVFKERIWGGTALADFGYTI~~S~~QRTGEC~~W~~AAFAAHQNGQS~~V~~VQ--NGMYKGFTL
SE

10 20 30 40 50 60

190 200 210 220 230 240
papa_carpa.p LLDCDRRSYGCNGG--YPWSALQLVAQYGIHYRNT~~P~~YEGVQ~~R~~YCRSREKGPYAAKTD
GV

YJDE LWEHHRH~~L~~FGQLEGDRFPLLTKILDADQDLSVQ-VHPND---EYANIHEN~~E~~LG-KTE
CW

70 80 90 100 110

250 260 270 280 290 300
papa_carpa.p RQVQPYNEGALLY---SIANQPVS~~V~~LEAAGKDFQLYR----GGIFVG~~G~~PCGNKVDHA
VA

YJDE YIIDCQKDAEIIYGHNATT~~K~~ELTTMIERGEWDELLRRVKVKPGDFFYVPSGT----

VH 120 130 140 150 160 170

300 310 320 330 340
papa_carpa.p AVGYGP~~N~~YIL~~H~~NSW~~S~~TGW~~G~~EN~~Y~~IRIKRG~~T~~GNSYGV~~C~~GLYT~~S~~F~~Y~~PVKN

YJDE AIGKG~~I~~LAET~~I~~NSITTYR~~L~~YDRKDAEGKL~~R~~ELHLKKSIEVIEVPSI~~P~~ERHTVHH
EQ

180 190 200 210 220 230

GC 381

Figure 2

7/7

SCORES Init1: 1141 Initn: 1252 Opt: 1256 z-score: 1441.1 E(): 0
 Smith-Waterman score: 1256; 56.6% identity in 316 aa overlap

59 yjde.pep FT PMI KT 60	10 20 30 40 50 MTTEPLFFKPVFKERIWGGTALAD-FGYTIPSQRTGECWAFAAHQNGQSVVQNGMYKG : : : : : : : : : : : MTQSPIFLTPVFKERIWGGTALRDRFGYSIPSESTGECWAISAHPKGPSTVANGPYKG ↑ 40 50
19 yjde.pep YI PMI YI 20	60 70 80 90 100 110 1 LSELWEHHRHLFGQLEGDRFPPLTKILDADQDLSVQVHPNDEYANIHENGEGLGKTECW : : : : : : : : : : : : LIELWEEHREVFGGVEGDRFPPLTKLLDVKEDTSIKVHPDDYYAGENEEGELGKTECW 70 80 90 100 110 1
79 yjde.pep IL PMI AL 80	120 130 140 150 160 170 1 IDCQKDAEIIYGHNATTKEELTTMIERGEWDELLRRVKVKPGDFYYVPSGTVAIGKG : : : : : : : : : : IDCCKENAEIIYGHHTARSKTELVTMINSGDWEGLLRRIKIKPGDFYYVPSGLHALCKG ↑ 130 140 150 160 170 1
39 yjde.pep TT PMI IK 40	180 190 200 210 220 230 2 ALETQQNSDTTYRLYDVRKDAEGKLRELHLKKSIEVIEVPSI PERHTVHHEQIEDLL : : : : : : : : : VLETQQNSDATYRYYDVRLDNSGSPRELHFAKAVNAATVPHVDGYIDESTESRKGIT ↑ 190 200 210 220 230 2
99 yjde.pep GL : PMI QM 00	240 250 260 270 280 290 2 TLIECAYFSVGKWNLSGSASLHQKQKPFLLISVIEGEGRMISGEYVYPFKKGDHMLLPY : : : : : : : TTVQGEYFSVYKWDINGEAEQDESFLICSVIEGSGLLKYEDKTCPLKKGDHFILPA 250 260 270 280 290 3
yjde.pep PMI	300 310 GEFKLEGYAECLVSHL : : PDFTIKGTCTLIVSHL 310

GC 381 Figure 3

SCORES Init1 1128 Initn: 1128 Opt: 1 z-score: 1418.2 E():

0

Smith-Waterman score: 1236; 55.6% identity in 313 aa overlap

		10	20	30	40	50	
59	yjde.pep	MTTEPLFFKPVFKERIWGGTALAD-FGYTIPSQRTGECWAFAAHQNGQSVVQNGMYKG					
FT		: : : : : : : : : : :					
YDHS		MTHPLFLEPVFKERLWGGTKLRDAFGYAIPSQKTGECWAWSAHAHGSSSVKNGPLAG					
KT		↑					
19	yjde.pep	LSELWEHHRHLFGQLEGDRFPPLTKILDADQDLSVQVHPNDEYANIHENGEGLKTECW					
YI		: : : : : : : : : : : : : : : : : : : :					
YDHS		LDQVWKDHPEIFGFPDGKVFPLLVKLLDANMDLSVQVHPDDDYAKLHENGDLGKTECW					
YI		60 70 80 90 100 110 1					
79	yjde.pep	IDCQKDAEIIYGHNATTKEELTTMIERGEWDELLRRVKVKPGDFFYVPSGTVAIGKG					
IL		: : : : : : : : : : : : : : :					
YDHS		IDCKDDAELILGHASTKEEFKQRIESGDWNGLLRRIKIKPGDFFYVPSGTLHALCKG					
TL		↑↑					
120		130 140 150 160 170 1					
39	yjde.pep	ALETQQNSDTTYRLYDYDRKDAEGKLRELHLKKSIEVIEVPSIPERHTVHEQIEDLL					
TT		↓↓					
YDHS		VLEIQQNSDTTYRVYDYDRCNDQGQKRTLHIEKAMEVITIPHIDKVHTPEVKEVGN					
II		↑↑					
180		190 200 210 220 230 2					
99	yjde.pep	TLIECAYFSVGKWNLSGSASLQQKPFLLISVIEGEGRMISGEYVYPFKGDHMLPY					
GL		: : : : : : : : : : : : : : : :					
YDHS		VYVQSDYFSVYKWKISGRAAFPSYQTYLLGSVLSGSGRIINNGIQYECNAGSHFILPA					
HF		↑↑					
240		250 260 270 280 290 2					
yjde.pep		300 310					
		GEFKLEGYAECLIVSHL					
YDHS		: : :					
		GEFTIEGTCEFMISHP					
300		310					

GC 381

Figure 4

10 30
atgacgcatttttaggcctgtcttaagaaagactatgg
M T H P L F L E P V F K E R L W

50 70 90
ggagggacgaagcttcgtgacgctttggctacgcaataccctcacaa
G G T K L R D A F G Y A I P S Q

110 130
aaaacaggtgagtgctggccgttctgcacatgcccattggctcg
K T G E C W A V S A H A H G S S

150 170 190
tctgtaaaaatggcccgctggcagaaagacacttgatcaagtatgg
S V K N G P L A G K T L D Q V W

210 230
aaagatcatccagagatattcggttccggatggtaagggtttccg
K D H P E I F G F P D G K V F P

250 270 2
ctgctggtaaagctgctggacgccaatatggatctccgtgcaagtc
L L V K L D A N M D L S V Q V

90 310 330
catcctgatgatgattatgcaaaactgcacgaaaatggcgacccgt
H P D D D Y A K L H E N G D L G

350 370
aaaacggagtgctggatatcattgattgcaaagatgacgccaacta
K T E C W Y I I D C K D D A E L

390 410 430
attttgggacatcatgcaaggacaaaagggagatgttcaaacaacgata
I L G H H A S T K E E F K Q R I

450 470
gaaagcggtgattggAACGGGCTGCTgaggcgaatcaaggcca
E S G D W N G L L R R I K I K P

490 510 5
ggagatTTTATGTGCCAAGCGGTACACTCCATGCTTATGTAAG
G D F F Y V P S G T L H A L C K

30 550 570
ggAACCCCTGCTTGAATCCAGCAAAACTCTGATAACACATATCGC
G T L V L E I Q Q N S D T T Y R

590 610
gtatacgattatgaccgctgtaatgaccaggccaaaaagaactctt
V Y D Y D R C N D Q G Q K R T L

630 650 670
catatagaaaaagccatggaaagtcatacgataaccgcatacgataaa
H I E K A M E V I T I P H I D K

GC 381 Figure 5A

gtgcatacaccggaagtaaaagaagttggtaacgctgagatcattgtt
V H T P E V K E V G N A E I I V

730 750 7
tatgtgcaatcagattattctcagtgtacaaatggaagattagcggc
Y V Q S D Y F S V Y K W K I S G

70 790 810
cgagctgctttccttcataatcaaacctattgctggggagtgttctg
R A A F P S Y Q T Y L L G S V L

830 850
agcggatcaggacgaatcataaataatggtattcagtagatgaatgcaat
S G S G R I I N N G I Q Y E C N

870 890 910
gcaggctcacactttattctgcctgcgcatttggagaattacaata
A G S H F I L P A H F G E F T I

930
gaaggaacatgtgaattcatgatatctcatcct
E G T C E F M I S H P

GC 381

Figure 5B

10 30
 atgacgcattaccgattttctaacgcctgtgtttaaagaaaaatc
 M T Q S P I F L T P V F K E K I

 50 70 90
 tggggcggaaaccgcattacgagatagattggatacagtattcctca
 W G G T A L R D R F G Y S I P S

 110 130
 gaatcaacggggaaatgctggccattccgctcatccaaaaggaccg
 E S T G E C W A I S A H P K G P

 150 170 190
 agcaactgttgcattggccgtataaaggaaagacattgatcgagctt
 S T V A N G P Y K G K T L I E L

 210 230
 tgggaagagcaccgtgaagtattcggccgttagaggggatcggttt
 W E E H R E V F G G V E G D R F

 250 270 2
 ccgcttcgtacaaagctgctggatgtgaaggaagatacgtaattaaa
 P L L T K L L D V K E D T S I K

 90 310 330
 gttcaccctgatgattactatgccggagaaaacgaagagggagaactc
 V H P D D Y Y A G E N E E G E L

 350 370
 ggcaagacggaatgctggtacattatcgactgttaaggaaaacgcagaa
 G K T E C W Y I I D C K E N A E

 390 410 430
 atcatttacgggcatacggccgctcaaaaaccgaacttgcataatg
 I I Y G H T A R S K T E L V T M

 450 470
 atcaacagcggtaactgggagggcctgctgcaagaatcaaaaattaaa
 I N S G D W E G L L R R I K I K

 490 510 5
 ccgggtgatattctattatgtgccgagcggaaacgcgtgcacgcattgtgc
 P G D F Y Y V P S G T L H A L C

 30 550 570
 aaggggggcccttggtagagactcagaaaaattcagatgccacatac
 K G A L V L E T Q Q N S D A T Y

 590 610
 cgggtgtacgattatgaccgtcttgcatacgaacggaaagtccgagagag
 R V Y D Y D R L D S N G S P R E

 630 650 670
 cttcatttgcatttgcatttgcatttgcatttgcatttgcatttgcatttgc
 L H F A K A V N A A T V P H V D

690 710

GC 381

Figure 6A

gggtatatacatgacatcgacagaatcaagaaaaggataaccattaaa
G Y I D E S T E S R K G I T I K

730 750 7
acatttgcacaggaaatattttcggttataaatgggacatcaat
T F V Q G E Y F S V Y K W D I N

70 790 810
ggcgaagctgaaatggctcaggatgaatccttctgattgcagcgtg
G E A E M A Q D E S F L I C S V

830. 850
atagaaggaaagcggttgctcaagtatgaggacaaaacatgtccgctc
I E G S G L L K Y E D K T C P L

870 890 910
aaaaaagggtgatcaacttatttgccggctcaaatgcccgtttacg
K K G D H F I L P A Q M P D F T

930
ataaaaggaaacttgacccttacgtgtctcatatt
I K G T C T L I V S H I

GC 381

Figure 6B

